AD_____

Award Number: W81XWH-08-2-0015

TITLE: Comparing Virtual Reality Exposure Therapy to Prolonged Exposure in the Treatment of Soldiers with PTSD

PRINCIPAL INVESTIGATOR: Dr. Gregory Gahm, PhD

CONTRACTING ORGANIZATION: The Geneva Foundation
Tacoma, WA 98402

REPORT DATE: June 2014

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT:

X Approved for public release; distribution unlimited

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
June 2014	Final	1 June 2008 – 31 May 2014

4. TITLE AND SUBTITLE
Comparing Virtual Reality Exposure Therapy to Prolonged Exposure in the Treatment of Soldiers with PTSD

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-08-2-0015

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Dr. Gregory Gahm, PI; Dr. Greg Reger, Co-PI
Dr. Albert Rizzo, AI; Dr. Patricia Koenen-Woods, AI; Dr. Nancy Skopp, AI

email:kzink@genevausa.org

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
The Geneva Foundation
917 Pacific Ave, Suite 600
Tacoma, WA 98402

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, MD 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public distribution; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
Prolonged exposure (PE) is one of the most researched psychotherapies for PTSD. Virtual reality exposure (VRE) has demonstrated growing support as an innovative method for activating the trauma memory during exposure. However, there is limited research on the effectiveness of either treatment with active duty military personnel and there are no head-to-head clinical trials. Soldiers with PTSD from deployments to Iraq or Afghanistan (N = 162) were randomized to 10 sessions of either PE or VRE or were assigned to a minimal attention wait list. All assessments were conducted by a psychologist blind to treatment group. External, independent treatment fidelity reviews were conducted for both treatments. Service members were assessed before randomization, after 5 sessions, at posttreatment, and 3- and 6-months posttreatment. PTSD was assessed with the Clinician Administered PTSD Scale (CAPS). Data were analyzed at the end of the performance period for this study but prior to the end of the period of performance for another grant/recruitment site. Results indicated significant improvement in symptoms for PE and VRE relative to the waitlist control. Contrary to expectations, PE demonstrated greater PTSD symptom reduction relative to VRE, which was statistically significant at the 12-week follow-up assessment. These findings extend previous findings on the efficacy of exposure therapy to an active duty military population and highlight important areas for future research of VRE.

15. SUBJECT TERMS
PTSD, virtual reality exposure therapy (VRET), prolonged exposure therapy (PE)

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			USAMRMC
U	U	U	UU	38	19b. TELEPHONE NUMBER *(include area code)*

Table of Contents

Page

INTRODUCTION.

The rationale for this study was based on growing evidence that demonstrated that Virtual Reality Exposure Therapy (VRET) was a high-quality, effective treatment for PTSD (1-2) with the potential to improve access to care for Soldiers who would otherwise avoid treatment (3). Although prolonged exposure (PE) is considered one of the most effective cognitive-behavioral therapy (CBT) for treatment of posttraumatic stress disorder (PTSD), there were no randomized, controlled trials among an active duty military population. Furthermore, there were reasons why it may not be the most viable option for many Soldiers. First, PE requires a level of emotional engagement in the re-living of the trauma that many Soldiers are unable to obtain (4). Second, stigma and concerns about how Soldiers will be perceived by peers and leadership has a dramatic impact on whether a Soldier will seek care (5). VRET may address these concerns and could theoretically improve treatment outcomes and access to care by augmenting the patient's re-living of the trauma with a sensory-rich environment (3) and moderating stigma perceptions by offering non-traditional treatment that may be a preferable option for many Soldiers who are reluctant to seek out traditional talk therapies. Despite its promise as a viable treatment option, few studies have examined VRET for combat-related PTSD and there are no published studies that have compared VRET to PE in the treatment of combat-related PTSD.

The purpose of this randomized, single blind study was to evaluate the efficacy of VRET by comparing it to PE and a waitlist (WL) group in the treatment of PTSD in active duty (AD) Soldiers with combat-related trauma. The study was designed to test the general hypotheses that 10 sessions of VRET would successfully treat PTSD, therapeutically affect levels of physiological arousal, and significantly reduce perceptions of stigma toward seeking behavioral health services. Soldiers returning from deployments to Iraq who were diagnosed with combat-related PTSD following administration of the Clinician-Administered PTSD Scale (CAPS) were randomized to one of three groups: 1) PE; 2) VRET; or 3) WL. Soldiers underwent clinical assessments at baseline and after 5 and 10 treatment sessions. Outcome measures were also collected at 12 and 26 weeks post-treatment. Physiological arousal, patient satisfaction with treatment, and stigma toward seeking behavioral health services were also explored.

KEYWORDS.

Virtual Reality Exposure Therapy (VRET)
Prolonged Exposure Therapy (PE)
Post-Traumatic Stress Disorder (PTSD)
Clinician-Administered PTSD Scale (CAPS)

BODY.

Overview

This study was a randomized, waitlist-controlled clinical trial in which post-Iraq, post-Afghanistan deployed Soldiers with deployment-related PTSD were randomized to one of three groups: 1) PE (n = 54), 2) VRET (n = 54), or 3) Waitlist (WL; n = 54).

The objectives/hypotheses of the VRET/PE study were as follows:

1. We will test the hypothesis that 10 sessions of VRET and PE will reduce PTSD symptoms compared to the waitlist.

2. We predict that 10 sessions of VRET will significantly reduce PTSD symptoms relative to PE and Waitlist assignment.
3. We will examine physiological responses during treatment to test the hypothesis that VRET will result in heightened in-session physiological responses compared to PE. In addition, we predict that VRET will result in greater reductions in physiological responses after 10 treatment sessions compared to PE.
4. We will determine whether Soldiers report reduced fears of treatment stigma following VRET compared to PE.
5. We predict that Soldiers completing 10 sessions of VRET will have higher levels of treatment adherence (lower dropout rates) and ratings of treatment satisfaction than Soldiers completing 10 sessions of PE.

Participants

All participants were diagnosed with current PTSD as assessed by the Clinician-Administered PTSD Scale (CAPS). The diagnosis of PTSD was made by a doctoral level psychologist. To ensure reliable diagnostic procedures, our consultant, Dr. Barbara Rothbaum, trained all psychologists in formalized CAPS assessment procedures. Additional participant inclusion criteria included: (a) history of deployment in support of OIF/OEF, and (b) a non-sexual assault, deployment-related trauma that met criteria for PTSD according to the CAPS. Participants also had to agree not to initiate other psychotherapy for PTSD or new psychotropic medications.

After returning home from a deployment, Soldiers commonly experience a period of psychological readjustment during which most return to baseline functioning without treatment. To ensure that any treatment effects observed in the proposed study were not due to the normal recovery process, we excluded Soldiers who experienced a trauma within the previous 3 months. Additional exclusion criteria included: (a) a history of schizophrenia, bipolar, or other psychotic disorder, (b) a history of organic brain disorder, (c) current suicidal risk or self-mutilating behavior, as indicated by hospitalization in the past 6-months for risk of self-harm (d) an ongoing threatening situation (e.g. domestic violence), (e) current drug or alcohol dependence, (f) a history of seizures (a risk factor for VR adverse events), (g) prior history of PE therapy for PTSD, (h) a physical condition that interfered with the proper use of the Virtual Reality head mounted display or its peripherals, or (i) a loss of consciousness for a duration of greater than 15 minutes since entering active duty military service. Participants must have been stable on medications for at least 30 days.

Recruitment

Participants were recruited from the Behavioral Health Service at Madigan Army Medical Center at Fort Lewis, WA. Recruitment of Soldiers for study participation began 05 May 2009 and ended in April 2013. The final follow-up data was collected in November 2013. A total of 485 soldiers were referred, 296 consented, and 162 soldiers were randomized. One hundred thirty four subjects screen failed. Of the 162 randomized subjects, 3 were deemed ineligible after randomization (but retained in the intent to treat analyses), 89 completed all study related assessments (47 of which were randomized to waitlist), and 70 subjects withdrew after completing some portion of study assessments, but dropped out prior to the final 6-month follow-up assessment.

Of the 70 participants who did not complete all study requirements, 38 participants dropped out during the treatment phase of the study, 11 subjects were lost to follow-up during the treatment phase of the study, 18 were lost to follow-up or withdrew during the follow-up portion of the

study, and 2 were withdrawn by the study team. Treatment Fidelity reviews were completed on 15% of all sessions recorded throughout the conduct of this study.

Prolonged Exposure Therapy Protocol

Prolonged Exposure therapy consisted of 10 treatment sessions (lasting 90-120 minutes each), delivered weekly or twice-weekly, although flexibility was allowed to accommodate Soldier's training schedules. The formal protocol for prolonged exposure was followed. In the initial two sessions, the patient and therapist discussed the treatment rationale, talked about the client's reactions to trauma, and collaboratively developed a hierarchy of anxiety-provoking situations for in vivo exposure homework assignments. Session 3 marked the first imaginal exposure session and subsequent discussion of the exposure experience. Sessions 4-9 focused on prolonged imaginal exposure during which Soldiers revisted the trauma in as much detail as possible in the present tense, with subsequent discussions of their thoughts and feelings. Subjective Units of Distress scale were gathered every 5 minutes during imaginal exposure. Homework assignments following sessions 3-9 included listening to taped imaginal exposure sessions and in vivo exposure assignments. The final session included a final imaginal exposure, discussion of in vivo exposure, and a treatment progress review. The final part of the session focused on follow-up assessments and the termination of treatment.

Virtual Reality Exposure Therapy Protocol

The VRET protocol followed the same procedures as the PE protocol with the primary exception that all instances of imaginal exposure were augmented by immersion into *Virtual Iraq* environments. Similar to procedures for imaginal exposure, Soldiers revisited their trauma, telling it in the first person, present tense while the therapist customized Virtual Iraq to resemble events described. Two *Virtual Iraq* environments were utilized, specifically a city and a convoy environment. The two environments provided the clinician with flexibility to determine which environment best matched the patient's needs, based on her or his combat-related experiences. Both environments could be adjusted to match time of day (dawn, day, dusk, night), weather condition (sunny or sandstorm), and relational viewpoint (e.g., driver or passenger seat) to best reconstruct the patients traumatic experience. As soldiers navigated through these environments, the clinician could activate different audio (i.e., incoming mortars, weapons fire, voices, wind, etc.) and audiovisual stimuli (e.g., helicopter flyovers) to further approximate the traumatic experience.

Protocol Adherence

All therapy sessions were video recorded and 15% of planned sessions were randomly selected in advance for independent rating of treatment adherence and competence. Therapists were unaware of which sessions would be sent out for adherence review. Coders were not involved in other aspects of the study and were selected for this role based on experience as investigators on previous clinical trials of PE (Mary Heekin) and VRE (Judith Cukor). Treatment adherence forms used in previous clinical trials of PE (Barbara Olasov Rothbaum, Astin, & Marsteller, 2005) were used for PE and adapted for VRE. Videos were coded, reviewed, and feedback provided to therapists on an on-going basis throughout the trial for fidelity review and adherence monitoring (Barber, Triffleman, & Marmar, 2007).

Outcome Measures

Screening

The following measures were administered prior to randomization to ensure eligibility and capture baseline data.

Clinical and Stigma Outcomes

The following clinical and stigma outcome measures were administered at baseline, and after 5 and I0 treatment sessions. The CAPS, PCL, IASMHS, and stigma measures were also assessed at 12 and 24-weeks following treatment.

I) *Clinician-Administered PTSD Scale (CAPS)* (6). The CAPS is a structured interview that assesses all DSM-IV PTSD criteria in terms of frequency and intensity. Scores are computed for Intrusion, Avoidance, and Hyperarousal symptom clusters, as well a Total score. The CAPS is commonly used as a primary outcome measure in PTSD clinical trials (7). The CAPS Current and Lifetime Version, which measures a one month symptom-duration, was used for the Baseline and Follow-up assessments. The CAPS One Week Version, which measures a one week symptom duration, was used to assess participants at baseline and after Treatment Sessions 5 and 10. PTSD severity as measured by CAPS served as the primary PTSD outcome in this study.

2) *PTSD Checklist (PCL)* (8). The PCL is a self-report measure that evaluates all 17 PTSD criteria using a 5-point Likert scale. Sensitivity and specificity are reportedly .82 and .83, respectively for detecting DSM PTSD diagnoses.

3) *Beck Depression Inventory-II (BDI-II)* (9). This self-report measure of depression contains 21-items that are rated on a 4-point scale..

4) *Beck Anxiety Inventory (BAI)* (10). The BAI is a self-report measure consisting of 21 items designed to discriminate anxiety from depression. It has high internal consistency (.92) and I-week test-retest reliability (.75) and discriminates anxious from nonanxious diagnostic groups.

5) *Inventory of Attitudes Toward Seeking Mental Health Services (IASMHS)* (11-12). The IASMHS is a 24 item assessment of help-seeking attitudes. It includes the following three factors based on components of Ajzen's Theory of Plarmed Behavior (13): Psychological Openness, Help-seeking Propensity, and Indifference to Stigma. Alpha coefficients for the subscales range from. 79 to .82, and internal consistency for the full inventory is .87. Test-retest reliability for the factors ranges from moderate to high. Convergent validity is demonstrated by effectively differentiating those who would and would not use services.

6) *Perceived Stigma Measures.* Stigma was measured using two 5-question assessment scales. 1) The 5-Item Perceived Stigma Scale was adapted from a scale developed by Komiya (14), and later adapted for use in a study of veterans by Pyne et al (84), who found that depression severity scores were associated with higher levels of perceived stigma. Komiya (14) reported a coeffieient alpha of 0.72. As with the Pyne study, questions were adapted to receiving help for PTSD. 2) The second measure is a scale adapted from an inventory concerning stigmatization associated with completing psychological assessments (15) (5).

In-Session Assessments

The following assessments were used to determine levels of emotional and physiological engagement during treatment sessions.

1) *Subjective Units of Distress (SUDs)* (16).Ranging from 1 to 100, Subjective Units of Distress were gathered every 5 minutes during imaginal exposure to determine levels of distress and engagement in the situation.

2) *Physiological Data.* Heart rate, skin conductance, respirations, and peripheral skin temperature data were collected with the Biopac MPl50 (Biopac Systems, Inc.). The final analyses of these data have not yet been completed and will not be summarized below.

Patient Satisfaction Measure

The Client Satisfaction Questionnaire (CSQ) is an 18-item self-report measure of general satisfaction with treatment. Participants were asked to rate variables on a 4-point scale including the kind of service, treatment staff, quality of service, amount, length and quantity of service, outcome of service, general satisfaction, and procedures. Internal consistency and construct validity have been established (17) and the measure is widely used in research.

Statistical Analyses

Treatment adherence. We compared the proportion of participants completing all 10 treatment sessions in the VRET group to the PE group using a two-sample difference of proportions test of the null hypothesis that the completion proportion of the VRET group would be less than or equal to that of the PE group. We also estimated Kaplan-Meier curves for a graphical assessment of the rate of dropout or loss to follow-up as a function of the number of treatment sessions. We used a Poisson regression with the number of treatment sessions as an exposure variable to test the null hypothesis that the rate of nonadherence in the VRET group will be less than or equal to that of the PE group.

Behavioral outcomes. The primary hypothesis stipulates that both the PE and the VRET group will have improved scores on behavioral outcome measures as compared to the WL group. Moreover, the VRET group will have improved scores as compared to the PE group. Given the valence of the primary and secondary outcome measures, lower scores will indicate superiority. As such, the null hypothesis for testing would stipulate that the mean score of the experimental group of interest is greater than or equal to the defined comparison group. To account for attrition, we used linear mixed effects regression models (Singer & Willett, 2003) to estimate the differences in means of the behavioral outcomes. All study participants who provided data at baseline were retained in the intent-to-treat models through full information maximum likelihood. We estimated a random coefficient for the intercept to account for individual variability in baseline outcome scores. Measurement occasions were treated categorically with baseline as the reference value. The parameter estimates of interest were the interaction terms between treatment group assignment and measurement occasion at mid treatment and post treatment. These estimates indicated the magnitude and direction of the difference in means between the study groups at the particular measurement occasion. We report the regression coefficients (unstandardized differences), standard errors, and one-tailed *p*-values. We also report the regression coefficients standardized to single subject standard deviation of the outcome measure, defined as the square root of the sum of the random intercept variance and the residual variance (Raudenbush & Bryk, 2001). The CAPS “last week” measure was also analyzed “per protocol” by restricting the model estimation to those study subjects who had completed all ten treatment sessions and provided data at the post treatment measurement occasion. A final model of the CAPS “last week” and “last month” included data from all available measurement occasions to look at differences between the VRET and PE groups at the 12- and 26-week follow-up times. All models were estimated in Stata 12.1 (StataCorp, 2013) using restricted maximum likelihood.

Missing data. A key assumption of the linear mixed effects regression model is that the data were generated under a missing at random (MAR) or a covariate dependent assumption. Prior to estimating these models, we used a generalized linear model with a logit link and a Binomial error distribution to examine the association between the likelihood of dropout and several determinants, including CAPS scores, treatment assignment, and demographic variables. The results suggested that participants with lower education and those who did not identify as non-Hispanic white were more likely to drop out of the study during the treatment phase. Dropout was not related to CAPS scores. All regression models included education and race to improve the estimation. As a sensitivity analysis, we estimated a random coefficient selection model (Enders, 2010) which is appropriate for data that are missing not at random (MNAR). We specified a linear growth curve model for the first three measurement occasions using the CAPS "last week". We allowed for differences in the intercept and slope values based on treatment group assignment, and treatment group assignment, the latent intercept, and the latent slope were all determinants of binary indicators of dropout at both the mid treatment and post treatment assessments. We report the coefficients and 95% confidence intervals from this model as well as the model-implied treatment difference at the post treatment assessment for each of the treatment group comparisons. We estimated the selection model using M*plus* 7 (Muthén & Muthén, 2012).

Treatment satisfaction. We used a two-sample Student's t-test to compare the means of the CSQ at post treatment between participants assigned to the VRET and PE groups. For this analysis, we only included study subjects who completed all 10 sessions of the assigned treatment.

Challenges

A summary of challenges throughout this project have been compiled for this report.

Year 1: There was a delay in hiring a third member to the study team. The Geneva Foundation actively recruited for the study's project director/psychologist position. Interviews were conducted and an offer was extended but declined by the potential candidate.

A delay in participant recruitment occurred between IRB approval of the study on 13 March 2009 and 05 May 2009 while the DoD Individual Investigator Agreements were executed by the National Center for Telehealth and Technology (T2) and WRMC. As a result of this delay in planned participant recruitment, a new recruitment plan was developed. The revised plan included print advertising, a poster/flyer campaign and a provider informational video. The study team also conducted informational meetings to providers at various MAMC clinics, in order to increase referrals of potential participants.

Year 2: A decrease in active participant recruitment occurred due to the staffing shortage experienced while hiring and training the psychologist/project director and the loss of a previously trained investigator.

The NEXUS-10 equipment used for physiological data collection needed to be replaced due to preliminary data showing heavy artifact and data collection interference, likely due to the use of Bluetooth connection in our facility. Other biofeedback systems were investigated for possible use on this trial.

Year 3: Challenges identified during this reporting period included subject recruitment and retention. Despite continuing PI and sub-I clinic updates around the installation, recruitment

remained slower than desired. New web resources such as websites linking subjects directly to recruitment information were developed and approved by the IRB.

With this reporting period covering the second year of enrollment and follow-up of participants into the study, a challenge regarding subject retention become apparent. The study team consulted with subject matter experts on this topic (Dr. Andy Leon), and identified a possible protocol amendment that would include adding an additional questionnaire to measure subject initial intent to complete the study, as well as intent to return to the next treatment session.

Year 4: Challenges previously identified continued during this reporting period and included subject recruitment and retention. Despite continuing PI and sub-I clinic updates around the installation, recruitment remained slower than desired. Retention in treatment groups was also problematic. We previously added the "Intent to Return" measure at each session to improve identification of barriers to care and problem solving. However, of the total enrolled sample who had the opportunity to complete study participation (to include 26-week follow-up), nearly 50% attrition was observed. Although this may not be surprising for a highly mobile active duty population, it was expected to negatively impact our observation of the persistence of treatment effects. The investigators explored options for reducing missing data, including the possibility of amending the protocol to include phone follow-up assessment of symptoms and voluntary coordination with Command to increase support for study participation.

Please note that previous clinical trials of exposure therapy have found an average dropout rate of 21% (Hembree et al., 2003), though more recent studies of Veteran patients with PTSD have reported higher dropout rates. For example, an observational study of Veterans treated with prolonged exposure in clinical practice at a VA Medical Center reported a 34% drop out rate *when drop out was defined as completing 6 sessions of PE* (Our study requires 10 sessions). Similarly, a large RCT of women Veterans receiving 10 sessions of prolonged exposure reported a 38% drop out rate (Schnurr et al., 2007). As a point of reference, a meta-analysis of 19 medication trials for PTSD (Van Etten & Taylor, 1998) reported an average dropout rate of 32%.

In this context, our dropout rate, although scientifically undesirable, may not be surprising. In addition to the challenges faced by all patients in similar studies, our patients had to contend with training exercises, PCS, ETS, finalization of medical boards, military retirements, etc. This study represents one of the first studies of treating active duty military personnel with deployment-related PTSD and we expect it to make a meaningful contribution to the scientific literature on the care of our Warriors.

Year 5: With the end of grant funding projected for 31MAY2013, a no-cost extension was submitted to USAMRMC in July 2012. Without timely approval of this extension, two key personnel (ie - the research coordinator and primary assessing clinician) accepted offers for other positions. Without confirmation of another extension year on the project, the receipt of new referrals was stopped in February 2013 and the last subject was enrolled in April 2013 to ensure grant-funded time to complete the treatment portion of the study.

Year 6: All remaining assessments were completed for the subjects in follow-up. At the conclusion of data collection, the data were analyzed in accordance with statement of work. Given the loss of equipoise, the results were discussed with the Institutional Review Board. As a separate grant had funded recruitment at another Army site (Fort Bragg), investigators were asked to share the findings with currently enrolled participants at that site, offer patients randomized to VRET the opportunity to switch to PE, and cease randomizing patients to VRE.

Consultants on the study did not agree with this decision and expressed concerns to the funding agency. An external independent statistical review was then conducted by the statistician at Madigan Army Medical Center, who concurred with the actions of investigators and the IRBs decision. The funding agency hired a second external, independent statistician who did not concur with the actions of the investigators. The reports of both statisticians are included as attachments.

KEY RESEARCH ACCOMPLISHMENTS.

- This study is one of the first randomized trials of PE with active duty military personnel and the first clinical trial comparing VRET to a standard of care.
- The study reached its target enrollment for this funded recruitment site
- Transparently reported methods and findings will make important contributions to our understanding of how to care for Warriors with deployment related PTSD.

REPORTABLE OUTCOMES.

Summary of Principal Findings

By post treatment, 42.59% of participants in the VRET group were lost to follow up or had withdrawn from the study compared to 40.74% of participants in the PE group (d = .02, SE = 0.09, $p_{d<0}$ = .577). Figure 1 displays the Kaplan-Meier curves for treatment retention for participants assigned to the VRET and PE groups. Both groups showed substantial attrition over the course of treatment with most occurring by mid treatment. The Poisson regression coefficient comparing VRET to PE was 0.09 (SE = 0.30; $p_{b<0}$ = .622). For the assessment of both proportion and rate of dropout, we observed little difference between the treatment groups and failed to reject the null hypothesis.

Figure 1.

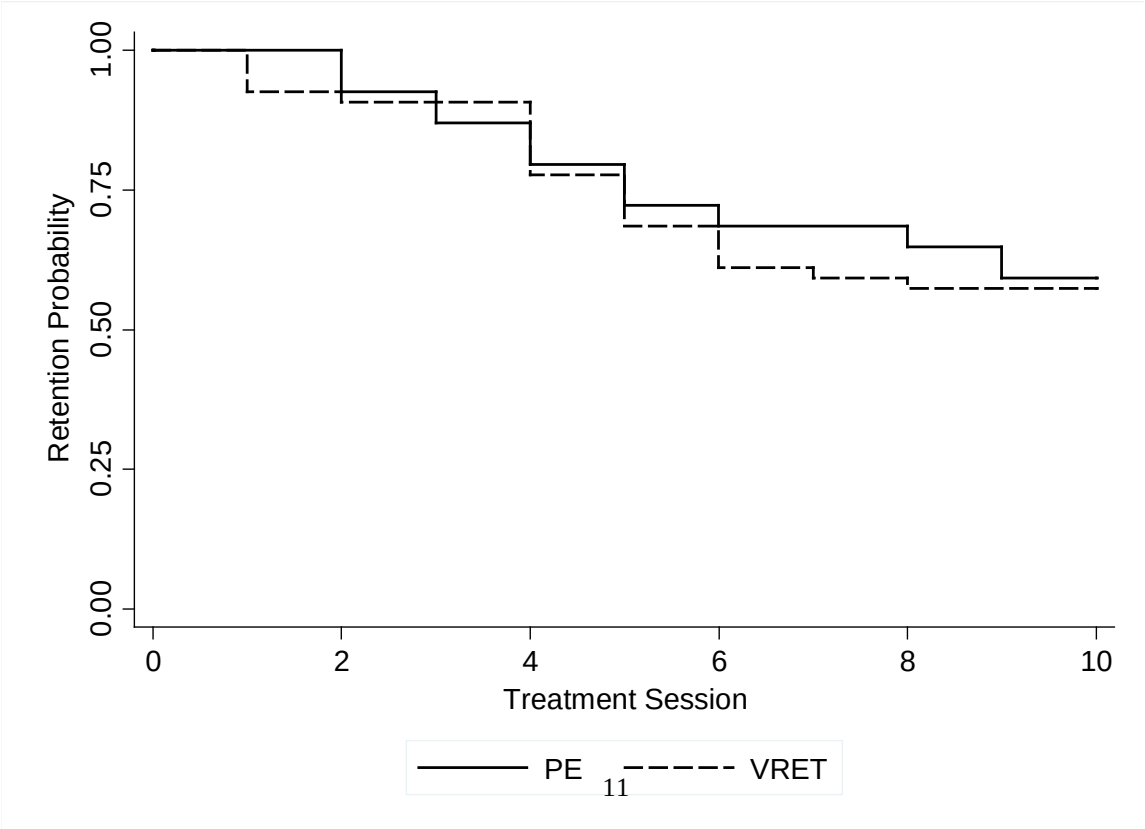


Table 2 provides descriptive data on the primary and secondary outcome measures for each treatment group at baseline, mid treatment, and post treatment. For the CAPS “last week” scores, the means decreased at each measurement occasion for all three study groups. The decreases were larger for the two active treatment groups. Internal consistency was high for all three study groups at all measurement occasions with the exception of baseline which was hindered by a compressed score range given the eligibility criteria for study participation. The secondary measures all showed adequate to good internal consistency reliability across measurement occasions and treatment groups.

Table 3 presents the results of the test of the primary hypothesis of superiority of the active treatments in reducing PTSD symptom severity over WL and of VRET over PE. Compared to participants in the WL, participants in PE had a decrease of 22.34 points on the CAPS “last week” and participants in VRET had a decrease of 13.30 points by post treatment. Both of these differences were statistically significant. The post hoc power to detect these differences was 1.00 for PE and 0.96 for VRET. For the comparison of VRET to PE, we observed a positive difference between the group means. This was consistent with the data in Table 1 that showed that the means post treatment were higher for those in the VRET group compared to PE. We failed to reject the null hypothesis of PTSD symptoms in the VRET group greater than or equal to those in the PE group at post treatment. The post hoc power to detect a one-tailed difference of a magnitude of 9.04 was 0.74 assuming it was in the anticipated direction of superiority. Given the direction favoring inferiority, our power was effectively 0.00. Increasing the sample size through additional randomization would not alter our ability to reject the null hypothesis. The results of these models, when restricted to treatment completers, were consistent with those observed from the intent-to-treat analysis (post treatment: PE – WL: b = -24.78, SE = 4.94, $p < .001$; VRET – WL: b = -12.63, SE = 5.00, $p = .006$; VRET – PE: b = 12.15, SE = 5.47, $p = .987$). The CAPS “last week” differences between VRET and PE were even greater at the 12-week (b = 15.07, SE = 6.03, $p_{b<0} = .006$) and 26-week (b = 13.91, SE = 6.70, $p_{b<0} = .019$) follow-up measurement times. The CAPS “last month” measure, which was only given at baseline and at the two post treatment follow-up assessments, was consistent with the CAPS “last week” at the follow-up measurement times (12-week: b = 16.98, SE = 6.30, $p_{b<0} = .996$ and 26-week: b = 14.42, SE = 6.99, $p_{b<0} = .980$).

Table 1. Means, minima, maxima, and internal consistency reliability for study measures, by treatment group and measurement time

Time	WL[1]				PE				VRET			
	n	M (SD)	Min., Max.	α	n	M (SD)	Min., Max.	α	n	M (SD)	Min., Max.	α
	CAPS (Week)											
Baseline	54	78.89 (16.87)	45, 114	0.70	54	78.28 (16.35)	54, 123	0.66	54	80.44 (16.23)	51, 111	0.66
Mid treatment	52	74.73 (21.78)	30, 117	0.83	39	65.03 (29.19)	11, 109	0.91	36	71.19 (23.27)	9, 115	0.86
Post treatment	47	68.06 (24.27)	10, 108	0.86	32	44.28 (33.73)	0, 121	0.94	30	57.07 (32.32)	0, 104	0.93
	PCL-C											
Baseline	54	60.30 (8.97)	33, 74	0.81	54	59.74 (9.09)	38, 79	0.79	54	61.85 (9.03)	41, 81	0.81
Mid treatment	52	55.58 (11.95)	31, 76	0.90	39	49.28 (14.85)	22, 80	0.94	36	53.17 (15.08)	20, 78	0.94
Post treatment	46	53.89 (11.77)	25, 78	0.88	32	40.63 (18.57)	17, 81	0.97	30	45.57 (15.88)	17, 69	0.95
	BDI-II											
Baseline	54	27.67 (9.99)	2, 52	0.89	54	28.02 (11.18)	10, 53	0.90	54	27.87 (9.19)	12, 51	0.86
Mid treatment	52	24.63 (10.70)	4, 50	0.91	39	21.69 (13.27)	1, 55	0.95	36	22.81 (11.44)	0, 45	0.92
Post treatment	46	25.63 (12.87)	2, 57	0.94	32	17.06 (16.18)	0, 59	0.97	30	18.50 (12.70)	1, 46	0.95
	BAI											
Baseline	54	23.81 (11.09)	2, 50	0.90	54	22.11 (9.34)	2, 42	0.86	54	24.57 (11.19)	8, 61	0.90
Mid treatment	52	21.35 (12.80)	0, 48	0.94	39	17.41 (9.72)	0, 40	0.89	36	19.78 (11.86)	3, 46	0.93
Post treatment	47	18.83 (11.93)	0, 49	0.93	32	13.28 (12.11)	0, 43	0.95	30	17.17 (12.80)	0, 50	0.94
	SSRPH											
Baseline	54	7.48 (3.10)	0, 15	0.83	54	7.04 (3.53)	0, 15	0.86	54	6.83 (3.52)	0, 14	0.84
Mid treatment	52	6.73 (3.35)	0, 15	0.84	39	5.59 (3.54)	0, 12	0.87	36	5.86 (2.94)	0, 11	0.77
Post treatment	47	6.77 (3.40)	0, 15	0.85	32	5.16 (3.73)	0, 14	0.89	30	5.43 (3.23)	0, 11	0.87
	IASMHS											
Baseline	54	43.91 (13.75)	18, 79	0.84	54	41.09 (15.93)	13, 79	0.89	54	43.80 (12.65)	19, 71	0.81
Mid treatment	52	45.48 (16.08)	13, 76	0.90	39	36.97 (16.10)	13, 76	0.91	36	41.36 (12.61)	12, 69	0.83
Post treatment	47	43.49 (15.43)	14, 70	0.90	32	34.56 (15.66)	9, 66	0.91	29	39.21 (13.80)	10, 63	0.87
	BASIS 24											
Baseline	54	1.54 (0.41)	0.5, 2.4	0.79	54	1.55 (0.46)	0.8, 2.7	0.83	54	1.60 (0.40)	0.6, 2.7	0.79
Mid treatment	52	1.45 (0.43)	0.4, 2.5	0.83	39	1.25 (0.57)	0.3, 2.5	0.91	36	1.32 (0.46)	0.3, 2.2	0.85

Post treatment	46	1.46 (0.51)	0.2, 2.7	0.87	32	1.03 (0.73)	0.0, 2.8	0.95	30	1.23 (0.60)	0.1, 2.5	0.90	

Note: WL = waitlist control; PE = prolonged exposure; VRET = virtual reality exposure therapy; M = mean; SD = standard deviation; α = Cronbach's alpha; CAPS = Clinician Administered PTSD Scale for DSM-IV; PCL-C = PTSD Checklist; BDI-II = Beck Depression Inventory – II; BAI = Beck Anxiety Inventory; SSRPH = Stigma Scale for Receiving Psychological Help; IASMHS = Inventory of Attitudes toward Seeking Mental Health Services; BASIS 24 = Behavior and Symptom Identification Scale; CSQ = Client Satisfaction Questionnaire.

[1]WL participants only completed measures up to post treatment.

Table 2. Intent-to-treat differences on primary and secondary outcome measures between the study groups at mid and post treatment assessments

Measure	PE – WL b (SE)	p^1	B	VRET – WL b (SE)	p^1	B	VRET – PE b (SE)	p^1	B
CAPS (Week)									
Midpoint	-9.32 (4.43)	.018	-0.40	-5.25 (4.52)	.122	-0.23	4.07 (4.77)	.803	0.18
Post treatment	-22.34 (4.69)	<.001	-0.97	-13.30 (4.77)	.003	-0.58	9.04 (5.11)	.961	0.39
PCL-C									
Midpoint	-5.29 (2.10)	.006	-0.43	-4.05 (2.14)	.029	-0.33	1.24 (2.27)	.708	0.10
Post treatment	-11.88 (2.23)	<.001	-0.96	-11.05 (2.27)	<.001	-0.89	0.83 (2.43)	.633	0.07
BDI-II									
Midpoint	-3.21 (1.74)	.032	-0.27	-2.32 (1.78)	.096	-0.20	0.90 (1.88)	.683	0.08
Post treatment	-8.83 (1.85)	<.001	-0.75	-7.59 (1.88)	<.001	-0.65	1.24 (2.02)	.723	0.11
BAI									
Midpoint	-2.14 (1.78)	.115	-0.19	-3.01 (1.81)	.048	-0.26	-0.87 (1.92)	.324	-0.08
Post treatment	-5.22 (1.88)	.003	-0.46	-4.74 (1.92)	.007	-0.42	0.48 (2.06)	.592	0.04
PSS									
Midpoint	-0.67 (0.60	.131	-0.20	-0.04 (0.61)	.474	-0.01	0.63 (0.65)	.836	0.19
Post treatment	-1.30 (0.63)	.020	-0.38	-0.72 (0.65)	.133	-0.21	0.58 (0.69)	.800	0.17
IASMHS									
Midpoint	-5.38 (2.19)	.007	-0.37	-3.58 (2.24)	.055	-0.24	1.79 (2.37)	.775	0.12
Post treatment	-6.55 (2.33)	.002	-0.45	-5.33 (2.39)	.013	-0.36	1.22 (2.57)	.683	0.08
BASIS-24									
Midpoint	-0.22 (0.07)	.002	-0.43	-0.18 (0.08)	.009	-0.36	0.03 (0.08)	.670	0.07
Post treatment	-0.46 (0.08)	<.001	-0.92	-0.31 (0.08)	<.001	-0.62	0.15 (0.09)	.957	0.30

Note: WL = waitlist control; PE = prolonged exposure; VRET = virtual reality exposure therapy; b = unstandardized coefficient; CI = confidence interval; B = standardized coefficient; CAPS = Clinician Administered PTSD Scale for DSM-IV; PCL-C = PTSD Checklist; BDI-II = Beck Depression Inventory – II; BAI = Beck Anxiety Inventory; SSRPH = Stigma Scale for Receiving Psychological Help; IASMHS = Inventory of Attitudes toward Seeking Mental Health Services; BASIS = Behavior and Symptom Identification Scale.
[1]One-tailed *p*-value to test the null hypothesis of treatment differences greater than or equal to zero in comparison to WL or PE.

The secondary outcome measures in Table 2 demonstrated similar patterns to the CAPS "last week" for the WL comparisons, with both VRET and PE demonstrating superiority over WL at post treatment on almost all measures. Specific to mental health treatment stigma, the means on the SSRPH were lower for both active treatment groups compared to WL, but these differences were statistically significant only for WL. We observed statistically significant reductions in the IASMHS for both treatment groups compared to WL. Similar to the CAPS "last week" comparisons, we observed small differences between the VRET and PE groups that favored PE. We failed to reject the null hypothesis that mental health treatment stigma in the VRET group would be greater than or equal to that of the PE group. Finally, participants in both the VRET and PE groups had high treatment satisfaction at post treatment (VRET: M = 3.47, SD = 0.47; PE: M = 3.52, SD = 0.52). The difference in means was trivial (d = -0.05, SD = 0.13, $p_{d>0}$ = .650) and we failed to reject the null hypothesis of satisfaction in the VRET group being less than or equal to that of the PE group.

Table 3 presents the parameter estimates from the random coefficient selection model of changes in the CAPS "last week". Noteworthy in these results was the lack of an association between the intercept and slope parameters with the indicators for drop out. This suggested that initial CAPS severity and the change in CAPS severity over time had little influence on the likelihood of dropping out of the study, consistent with the assumption of data missing at random employed in the models reported above. The model-implied differences at post-treatment between the three study groups were consistent with those reported in Table 2.

Table 3. Parameter estimates from a random coefficient selection model of participant dropout from baseline to post treatment and the model-based mean difference at post assessment

Parameter	b	95% CI	B
Intercept	78.99	74.47, 83.51	
PE	-0.44	-6.76, 5.88	-0.03
VRET	1.88	-4.33, 8.09	0.11
Slope	-4.21	-7.07, -1.34	
PE	-10.05	-16.37, -3.74	-0.57
VRET	-6.43	-12.35, -0.52	-0.37
Dropout at mid treatment[1]			
Intercept	-0.00	-0.05, 0.04	
Slope	0.01	-0.06, 0.07	
PE	2.38	0.66, 4.09	
VRET	2.61	0.96, 4.27	
Dropout at post treatment[1]			
Intercept	0.00	-0.04, 0.05	
Slope	0.06	-0.02, 0.15	
PE	1.31	-0.06, 2.67	
VRET	1.03	-0.28, 2.34	
Mean difference at post assessment			
PE-WL	-20.54	-34.59, -6.49	-1.16
VRET-WL	-10.99	-24.02, 2.04	-0.62
VRET-PE	9.55	-5.85, 24.95	0.54

Note: b = unstandardized coefficient; CI = confidence interval; B = standardized coefficient; PE = prolonged exposure; VRET = virtual reality exposure therapy.

[1]Regression of dropout on other model variables used a logit link and a Binomial error distribution.

CONCLUSION.

This study represents the first assessor blinded, randomized study of PE with active duty military members in addition to being the first randomized, controlled trial comparing PE and VRET. As such, its findings documenting the efficacy of these treatments represent a significant contribution to our understanding of effective treatments for PTSD. In addition, this study demonstrated that PE without VR was superior to PE with VR (VRET). This finding was in contrast to the hypothesized outcome that VRET would be superior to PE alone. While multiple potential explanations for this finding exist, including the potential for VR to be effective with some subgroups of patients, the finding stands.

REFERENCES.

1. Difede J, Cukor J, Patt I, Giosan C, Hoffman H: The application of virtual reality to the treatment of PTSD following the WTC attack. Ann N Y Acad Sci 2006; 1071:500-1
2. Rothbaum BO, Hodges LF, Ready D, Graap K, Alarcon RD: Virtual reality exposure therapy for Vietnam veterans with posttraumatic stress disorder. J Clin Psychiatry 2001; 62(8):617-22
3. Rothbaum BO, Hodges LF: The use of virtual reality exposure in the treatment of anxiety disorders. Behav Modif 1999; 23(4):507-25
4. Difede J, Cukor J, Ivy P, Giosan C, Hoffman H: The Application of Virtual Reality to the Treatment of PTSD Following the WTC Attack, in Psychobiology of Posttraumatic Stress Disorders: A Decade of Progress, vol 1071. Edited by Yehuda R. Malden, MA, Blackwell Publishing, 2006, pp 500-501
5. Hoge CW, Castro CA, Messer SC, McGurk D, Cotting DI, Koffman RL: Combat duty in Iraq and Afghanistan, mental health problems, and barriers to care. N Engl J Med 2004; 351(1):13-22
6. Blake DD, Weathers FW, Nagy LM, Kaloupek DG, Gusman FD, Charney DS, Keane TM: The development of a Clinician-Administered PTSD Scale. J Trauma Stress 1995; 8(1):75-90
7. Schnurr PP, Friedman MJ, Foy DW, Shea MT, Hsieh FY, Lavori PW, Glynn SM, Wattenberg M, Bernardy NC: Randomized trial of trauma-focused group therapy for posttraumatic stress disorder: results from a department of veterans affairs cooperative study. Arch Gen Psychiatry 2003; 60(5):481-9
8. Blanchard EB, Jones-Alexander J, Buckley TC, Forneris CA: Psychometric properties of the PTSD Checklist (PCL). Behav Res Ther 1996; 34(8):669-73
9. Beck AT, Steer RA, Ball R, Ranieri W: Comparison of Beck Depression Inventories -IA and -II in psychiatric outpatients. J Pers Assess 1996; 67(3):588-97
10. Beck AT, Epstein N, Brown G, Steer RA: An inventory for measuring clinical anxiety: psychometric properties. J Consult Clin Psychol 1988; 56(6):893-7
11. Mackenzie CS, Knox VJ, Gekoski WL, Macaulay HL: An adaptation and extension of the Attitudes Toward Seeking Professional Psychological Help scale. J of Applied Social Psych 2004; 34(11):2410-2435

12. Mackenzie CS, Gekoski WL, Knox VJ: Age, gender, and the underutilization of mental health services: the influence of help-seeking attitudes. Aging Ment Health 2006; 10(6):574-82
13. Ajzen I: From intentions to actions: A theory of planned behavior., in Action-Control: From Cognition to Behavior. Edited by Kuhl J, Beckman J. Heidelberg, Springer, 1985, pp 11-39
14. Komiya N, Good GE, Sherrod NB: Emotional openness as a predictor of college students' attitudes toward seeking psychological help. J Counsel Psychol 2000; 47(1):138-143
15. Britt TW: The stigma of psychological problems in a work environment: Evidence from the screening of Service Members returning from Bosnia. J of Applied Psych 2000; 30(8):1599-1618
16. Foa EB, Hembree E, Rothbaum BO: Prolonged Exposure Therapy for PTSD: Emotional Processing of Traumatic Experiences. New York, Oxford University Press, 2007
17. Larsen DL, Attkisson CC, Hargreaves WA, Nguyen TD: Assessment of client/patient satisfaction: development of a general scale. Eval Program Plann 1979; 2(3):197-207

APPENDICIES.

Appendix 1: A Statistical Review on the Justification of the Early Termination of the VRPE Study by Raywin R. Huang, PhD

Appendix 2: Independent Statistical Trial and Analysis Review (Nicole Close, PhD)

Appendix 3: Preliminary Results from a Randomized Controlled Trial of Prolonged Exposure and Virtual Reality Exposure Therapy for Soldiers with PTSD

PERSONNEL RECEIVING PAY FROM THE RESEARCH EFFORT:
AST, ERIKA
BALES, GARY
BARTLETT, LAUREN
FANTELLI, EMILY
KOENEN-WOODS, PATRICIA JO
MCCANN, RUSSELL
MUNROE, CHARNETTE
O'BRIEN, KAREN
PRUITT, LARRY
SHAW, AMBER
THOMAS, ELISSA
ZETOCHA, KIMBERLEE

A Statistical Review on the Justification of the Early Termination of the VRPE Study

COL (Dr) Andrew Wiesen

This document is about a statistical review for the justification to terminate the VRPE study that was prompted by an interim look at partially collected data.

Internal Review Board

Department of Clinical Investigation

Madigan Army Medical Center

4/14/2014

STATISTICAL REVIEW OF VRPE STUDY

A. Purpose of Review

To evaluate whether there are any justification for termination of study that is titled, "Comparing Virtual Reality Exposure Therapy to Prolonged Exposure in the Treatment of Soldiers with PTSD", after the study has taken an interim look that was not specified in the protocol-a protocol deviance.

B. Situation at Hand

Hitherto to this date 21st March 2014, the study has collected 162 subjects (54 subjects in each of the three arms, i.e. waitlist null condition (WL), prolonged exposure therapy (PE) and virtual enhanced therapy (VRET). The findings are presented in the report titled, "VRPE Study: Analysis Report", 21 March, 2014. These findings prompted the investigators to conclude with sufficient evidence that VRET (Virtual Reality Enhanced Therapy) treatment is inferior to PE (Prolonged Exposure Therapy) in terms of CAPS score (efficacy), and thereby, unnecessary to continue randomization between VRET and PE i.e. stopping the study.

C. Statistical Evaluation

The crux of the evaluation focuses on the primary hypothesis for which the sample size was based under section "Justification of sample size" (page 12).

The Primary Hypotheses

Hypothesis: *Virtual Reality Exposure Therapy (VRET) will significantly reduce PTSD symptoms compared to Prolonged Exposure (PE) and Waitlist (WL) assignment. (underline is mine)*

The Statistical Hypothesis

H_O: VRET $\geq$ PE & WL
H_A: VRET $<$ PE & WL

The Primary Measured Outcome Variable: Change in CAPS (at weeks 12 & 26) from baseline.

The Sample size Justification

"The results of a power analysis incorporating a conservative *effect size (f^2) of* ***0.09*** (estimated based on the literature cited above) and a Type-I error rate of 0.05 revealed that ***69*** subjects in each of three treatment groups would ensure adequate power to detect a true effect with 80% accuracy (power)."

What the Interim Analysis Showed

According to Table 4 of the analysis report, the mean and standard deviation for the week CAPS score (CAPS-W) for the three arms were as follows:

Table 4

Descriptive statistics of outcome measures for the VRPE study at each measurement occasion

	Waitlist			PE			VRET		
Time	n	*M*	*SD*	N	*M*	*SD*	n	*M*	*SD*
					CAPS-W				
Baseline	54	78.89	16.87	54	78.28	16.35	54	80.44	16.23
Mid treatment	52	74.73	21.78	39	65.03	29.19	36	71.19	23.27
Post treatment	47	68.06	24.27	32	44.28	33.73	30	57.07	32.32
12-week follow-up				27	36.63	31.80	26	55.88	31.10
26-week follow-up				24	38.33	28.49	17	54.47	28.62

Comment: Although 54 subjects were recruited for each of the three groups, there were loss of follow-up between baseline and 26 weeks, loss of 30 for PE and loss of 37 for VRET; loss of 7 between baseline and post-treatment for WL. The loss of subject has implication for loss of statistical power to capture "true" differences between groups for it is no longer the sample size as derived from above. Also there were quite a number of missing values.

The main outcome measure to be used for evaluation:

Since the primary hypothesis stated difference in reduction of PTSD symptoms as assessed by CRAPS scores, PTSD change scores were generated, i.e. (Post Treatment/FU-Baseline), and they were applied to determine difference between treatment groups for each of the post-treatment measures, i.e.at post-treatment, at 12 weeks, and 26 weeks. *However, this evaluation will only focus on the 26 week follow-up which is the end point for which the stop decision was made.*

Evaluation Strategy –Sensitivity Approach

1. **What is the statistical power for drawing a conclusion with the present data?**
2. **What would have been the statistical power if full set of data of 54 per group for drawing a conclusion?**
3. **What would have been the statistical power if missing data has been imputed for drawing a conclusion?**

1. **What is the statistical power for drawing the conclusion with the present data?**

Table 1

Means of PTSD-W Change Scores:

Treatment Group	**N**	**Mean**	**Std. Error**
PE	**24**	**-38.67**	**5.69**
WL	**47**	**-10.40**	**2.74**
VRET	**17**	**-23.58**	**6.65**

P=0.001

Note: Since WL do not have any assessments at 26 weeks, the assessments at post-treatment were carried forward to 26 weeks on the assumption that assessments were stable after post-treatment (depicted by Graph A).

Mean Change Differences Between Treatment Groups:

	PE	WL	VRET
PE	----	-28.26	-15.07
WL	<.001	----	13.18
VRET	0.133	0.147	----

Note: Upper diagonal are mean differences and lower diagonal are p-value generated by Bonferroni post-hoc c1omparison.

2. **What will be the statistical power for drawing the conclusion with full set of data of 54/group**

Table 2

Means of PTSD-W Change Scores:

Treatment Group	**N**	**Mean**	**Std. Error**
PE	**54**	**-38.67**	**3.79**
WL	**54**	**-10.40**	**2.55**
VRET	**54**	**-23.58**	**3.73**

P=0.001

Mean Change Differences Between Treatment Groups:

	PE	WL	VRET
PE	----	-28.26	-15.07
WL	<.001	----	-13.18
VRET	0..014	0.306	----

3. What will be the statistical power for drawing the conclusion if missing data has been populated? (Missing data populated by LOCF-If week 26 missing, it will be carried forward by week 12 or by post-treatment)

Table 3

Means of PTSD-W Change Scores:

Treatment Group	**N***	**Mean**	**Std. Error**
PE	**31**	**-41.67**	**5.56**
WL	**47**	**-10.40**	**2.74**
VRET	**30**	**-23.50**	**5.10**

P=0.001

*n differ by groups due to some missing data cannot be imputed, e.g. subjects only have baseline scores.

Mean Change Differences Between Treatment Groups:

	PE	WL	VRET
PE	----	-31.27	-16.41
WL	<.001	----	14.86
VRET	0..039	0.041	----

Summary:

The Effect Size, p-value and Statistical Power under the three conditions:

Situation	Effect Size[+]	p-value*	Statistical Power*
With the present results of sample size 17 & 21	0.263	0.133	53%
If full sample size of 54 per group had been collected**	0.263	0.014	87%
If missing data has been populated	-0.294	0.039	68%
*Generated by one-tailed test ** With the assumption that the mean and standard deviation remain invariant but the standard error would change due to increase sample size + Cohen's d=0.545; The selected effect size for sample size justification was 0.09;			

COMMENTS:
For the present condition PE differed insignificantly (p=0.133) from VRET by 15 points of increased reduction. If the study would have full 54 subjects per group, PE would have differed significantly (p<0.001) by the same magnitude. If missing data had been imputed, PE would have 16 points of significantly increased reduction (p=0.039).

However, the direction of this result under three conditions is counter to the primary hypothesis that VRET would be more efficacious in reducing CAPS week score than PE. Instead, PE was shown to be more efficacious than VRET with projected increase number of subjects, and when missing values were imputed under worse conditions (i.e. the lower scores were imputed for missing). *Hence, the projection of any further collection of subjects would prove the opposite, and it would be a futile effort in time and resources to continue the study.*

RECOMMENDATIONS:

This statistical evaluation suggests that continuation of the study will not change the results that point to the opposite direction of the primary hypothesis. Hence fore, this study is recommended to be terminated.

Submitted by:

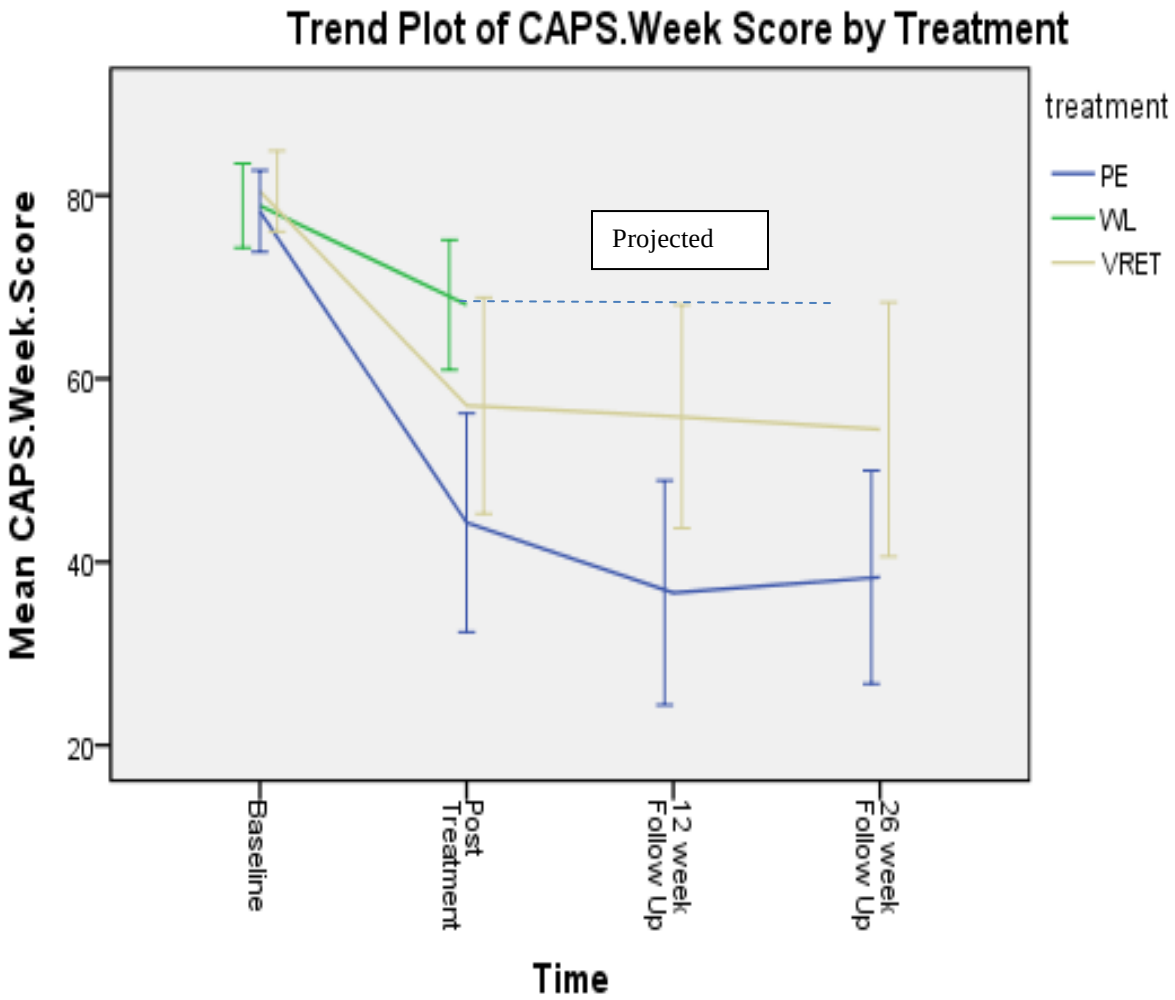
Raywin R Huang, Ph.D.

Senior Biostatistician & Chief

Biostatistics and Bioinformatics Service

Department of Clinical Investigation

Graph A



WL=WaistList;PE=Prolonged Exposure Therapy

VRET=Virtual Reality Enhanced Therapy

Error bars: +/- 2 SE

01 July 2014
Nicole C. Close, PhD
Independent Statistical Trial and Analysis Review

Materials Received:
1. Madigan Army Medical Center Clinical Investigator Protocol "Comparing Virtual Reality Exposure Therapy to Prolonged Exposure in the Treatment of Soldiers with PTSD"
2. 3 page summary of analysis for "Comparing Virtual Reality Exposure Therapy (VRET) to Prolonged Exposure (PE) in the Treatment of Soldiers with PTSD"

Verbal comments included that:

- A second site at Ft. Bragg was added to the study.
- The study was stopped early.
- The trial had an interim analysis.

Trial and Statistical Review:

1. *Assessment of Trial Assay Sensitivity*: The property of a clinical trial that is defined as the ability to distinguish an effective treatment from a less effective treatment. Without assay sensitivity, a trial is not internally valid and is not capable of comparing the efficacy of two or more interventions.

a. Hypotheses: The hypotheses presented in the protocol are NOT the same as the hypotheses stated in the summary of analysis:

Protocol: (verbatim)

Hypothesis: *(if applicable)* There are several hypotheses for this project:

1) Virtual Reality Exposure Therapy (VRET) will significantly reduce PTSD symptoms compared to Prolonged Exposure (PE) and Waitlist (WL) assignment.
2) VRET will result in heightened in-session physiological responses compared to PE. In addition, we hypothesize that VRET will result in greater reductions in physiological responses at treatment completion compared to PE.
3) Soldiers will report significantly reduced fears of treatment stigma following VRET compared to PE.
4) Soldiers completing VRET will have higher levels of treatment adherence (lower dropout rates) and ratings of treatment satisfaction than Soldiers completing PE.

Summary Analysis (verbatim)

The two primary hypotheses are of the VR/PE study were:

1. Psychotherapy that used prolonged or virtual reality-enhanced prolonged exposure (VR) would reduce the clinical symptoms of PTSD to a greater extent that a waitlist condition.

2. VR therapy would have greater efficacy in PTSD symptom reduction as compared to PE therapy.

Hypothesis Points to Consider:

- The protocol does not indicate if there is one primary hypothesis, if all three were primary hypotheses, or if any are secondary hypotheses.
- There is no indication which hypothesis(es) were considered for the sample size and power calculations and how the relationship between the 4 hypotheses were considered for power of the study.
- The analysis summary states ***different*** hypotheses from the protocol.

a. Compare VR to PE **and** VR to WL for reducing PTSD symptoms (protocol Hyp#1)
b. Compare VR to PE for heightened in-session physiological responses. (protocol Hyp#2a)
c. Compare VR to PE for reductions in physiological responses (protocol Hyp#2b)
d. Compare VR to PE for reduced fears of treatment stigma (protocol Hyp#3)
e. Compare VR to PE for higher levels of treatment adherence (protocol Hyp#4a)
f. Compare VR to PE for ratings of treatment satisfaction (protocol Hyp#4b)

a. Compare PE to WL **or** VR to WL for reducing PTSD symptoms (analysis Hyp#1a);
b. Compare VR to PE for PTSD symptoms reduction. (analysis Hyp#2)

- There is no indication about control of the alpha level for testing multiple hypotheses. Was a hierarchical hypothesis test used? Was an adjustment made for each hypothesis test?
- There is no discussion of the endpoint and timepoint for comparison for each of the analyses.

Sample Size and Power:

- The investigators chose a quantitative effect size (medium), rather than estimating a qualitative effect size to use for sample size and power calculation. This may be accepted in social research, but when a medium effect size is chosen, give that this method uses an assumed population parameter of effect sizes, you will choose the same n regardless of the accuracy or reliability of the instrument, or the narrowness of the diversity of the subjects.

- There is no discussion of the hypothesis used or if all were considered in choosing the appropriate sample size for the study and if the same statistical assumptions were used for each of the hypotheses and endpoints.
- The total sample size for the study was 54 subjects per treatment group (N=162). The analysis summary indicates that the all 162 participants were included, so it is unclear why it would be stated that this study was terminated early. The protocol update indicates that the screening number would increase to 300 so that 54 subjects could be recruited and randomized to each group.

Randomization:

- Other than stating that a random number generator was used, there is no statement or discussion of the randomization methodology.
    - This has a direct impact especially if the study was stopped early.
    - This has a direct impact on the analysis to be used for the study.
- It is important to understand what fixed allocation scheme was used (for example, simple randomization, blocked randomization, random permuted block).
- Was there stratification used for the study (by site (for multiple sites), by any other factor)? This is also important for the analysis of the study since any stratification factors must be considered in the analysis.
- Subjects are randomized to a treatment condition to being within a week. There is no statement in the analysis summary if any subjects dropped out between randomization and the first treatment condition "applied."

Analysis Population(s):

- The protocol indicates that an intent to treat analysis population and a completers analysis population would be used for all analyses. However, in the analysis summary there is no indication of the analysis population used.
- There is no definition of what constitutes a "completer." A formal statement of the definition should be included as well as the analysis.
- There is no CONSORT summary of those recruited, randomized and analyzed.

Missing Data:

- Neither the protocol nor the analysis summary indicate how missing data will be handled and how they were handled. The extent of missingness was not included in the analysis summary.
- Since Total Score from the CAPS was used as the primary analysis, a description of how that score was calculated was not given.

Statistical Analysis:

- The protocol indicates that CAPS Total Score is the primary endpoint for analysis. It is measured at Baseline, after treatment session 5, after treatment

session 10, at 12 weeks and 26 weeks. It is indicated that the WL group will not participate in the 12 and 26 week follow-up.

- ANCOVA proposed as the primary analysis to evaluate the effects of a course of VR or PE or Waitlist on PTSD symptoms; however, the analysis summary used a random-intercept linear regression model. There is no indication why the primary analysis methods were changed.
- Duncan's multiple range test was indicated for "all analyses" in the protocol to assess the significance of pair-wise comparisons. However, this test does not control for the family wise error rate at the nominal alpha level and any increase of power resulting from performing this test come from the intentional raising of the alpha levels and not a statistical improvement of the test.
- There is indication in the protocol that current therapy or other treatments that service members are receiving and will be analyzed as part of the treatment outcome measure. However, there is no indication of this inclusion in the analysis summary.

2. Analysis Summary Review:

    - It is assumed that the randomization allocation was 1:1:1, even though it was not stated.
    - Type of randomization implemented was not indicated.
    - There is no indication if randomization was stratified.
    - There is no indication if this was a single center or multi-center study.
    - There are two primary hypotheses stated in the summary analysis, but there is no indication that the statistical analyses were controlled for multiplicity. If not controlled, inflation of the Type I alpha should be discussed in relation to the findings for multiple primary hypothesis tests conducted.
    - The hypotheses indicated in the Analysis Summary would be more suited to testing with methods from the set of closed testing procedures that control the family-wise error rates.
    - There is no indication why the primary hypotheses were different from the protocol hypotheses.
    - The summary analysis should have a detailed listing of the number of subjects assessed for eligibility, randomized, allocated to each group, # received treatment, # loss to follow-up, # discontinued treatment, and the final number analyzed in the ITT group and the number analyzed for the Completers group.
    - There is no discussion of the primary variable of the CAPS total score and how this was calculated. This should be discussed in terms of missing data and outliers.
    - Baseline data must be shown in terms of demographics as well as the primary variable by treatment group.

- Analyses chosen for these hypotheses were not justified and were not controlled for multiple hypothesis testing.
- Interpretation of the analyses was not correct. For example, the hypotheses were stated as superiority hypothesis comparisons, but one was interpreted as an inferiority test. This was not an inferiority hypothesis.

If this trial was "terminated prematurely" there should have been a protocol amendment submitted to the IRB with the following items:

- Study Design update to include:
  - New sample size and power estimates
  - Inclusion of new sites
  - Randomization update and if stratification by site would be used or a shared central randomization schedule developed at the beginning of the study.
- A trial should only be terminated early if:
  - there were safety issues (adverse events (larger than expected or unanticipated);
  - An unbiased review of the data (blinded) has been conducted
- If any analyses are conducted for consideration of stopping a trial early, they should include both futility analyses and a summary of the conditional power.
- Design effects (either original or estimated from the current data) should be considered with the assumed true effect size and planned sample size and power.

None of this was discussed in the Summary of Principal Findings.

Findings:

1. Based on these two documents, the summary of principal findings does not follow the protocol.
2. The protocol has design flaws that may have implications on any statistical analyses.
3. There is no discussion of data collected, data quality, missingness, follow-up completeness in the protocol (how it will be handled) or the summary of principal findings (how they were handled).
4. It is highly unlikely that everything in the protocol was implemented as planned and there are no data presented for interpretation of that impact on the final analyses.
5. Statistical analyses presented do not follow the protocol planned analyses, and may not be the most robust and appropriate for this study design.
6. These data do not represent a planned or unplanned interim analysis nor should be interpreted as such based on the protocol.

Recommendations:

1. Any amendments to the protocol should be obtained and reviewed.
2. The raw data from the study should be obtained for review.
3. All analyses conducted should be obtained for review.
4. A complete report in regards to the trial design and implementation should be obtained to make a full determination of assay sensitivity.

Items to include are specifics about the:

- number of trial sites,
- final sample size and power calculations,
- qualitative effect size estimate and observed,
- randomization methodology,
- stratification factors,
- CONSORT diagram information,
- Baseline demographics by treatment group
- Descriptive statistics for the study
- Baseline and timepoint CAPS summary by Treatment Group
- Missing data report
- Loss to follow-up report
- Summary of data collection, management, edit checks and quality review
- Monitoring Report/Summary for the Study (source to database checks)

5. Implementation of a re-analysis of the data using appropriate statistical tests for the protocol primary hypothesis(es)
6. Post-hoc power analysis for sample size and effect size observed.
7. Appropriate interpretation based on correct statistical analyses.

08 August 2014

After a study team conference call to discuss review requirements and an additional one-on-one conference call with the study team data analyst, the additional study documents were obtained for review:

- Vrpe_anova.txt *(Primary analysis results conducted but not reported in the summary analysis)*
- Post-bragg.doc *(Protocol submission to IRB after introduction of the additional study site)*
- PreBragg.doc *(Protocol submission to IRB before introduction of the additional study site)*
- Version prior to data analysis.doc *(Protocol submission to IRB just prior to the reviewed data analysis)*
- AE Workbook updated from 2013 CR wwf.xlsx *(listing of Adverse Events)*
- Disposition_20140728.xlsx *(listing of missed visits and withdrawals)*
- Protocol Deviation_revised APR 2013 for CR.xlsx *(listing of reported protocol deviations for the study)*
- Posthoc_power_contrasts.rtf *(post hoc power contracts for two groups, timepoints)*
- Posthoc_power_rmanova_time1_3.rtf *(power analysis for the primary endpoint using 3 timepoints)*
- Posthoc_power-rmanova_time1_5.rtf *(power analysis for the primary endpoint using 5 timepoints)*
- Vrpe_data-20140723.xlsx *(extracted data from the study used as the primary endpoint for the study. These were not the raw data but the analysis endpoint created, no verification of the raw data for these endpoints.)*
- Vrpe_data_labels_20140723.xlsx *(data labels for the extracted data)*
- Vrpe_analysisrep_20140321_v2.docx *(described by the data analyst as a document to defend the statistical analyses performed and interpreted for the final study reports. Not a statistical analysis report document)*

Executive Summary:

Limited data were first available for review of this study and the subsequent statistical analysis presented. Reasons for early termination of the study, reasons for unblinding and analyzing the data and presentation of the data were not appropriate. Multiple protocols were produced that had been written and approved across a period of time with changes in the sample size and number of participating sites. The basic study methodology, design, endpoint and statistical analyses proposed for the primary and secondary endpoints remain unchanged across versions. There was a recalculation of the sample size based on a different effect size based on the literature cited by the study team.

The primary statistical analysis was conducted and presented for the study from one site and after 162 subjects were randomized (77.5% of the total sample size). This was not a planned interim analysis and the statistical analysis was not adjusted for this interim look at the data. There are multiple hypotheses and adjustment for multiple hypothesis testing was not conducted.

Table 4: Power Calculation

Treatment Groups (G) =	3 (PE, VRET, Waitlist)
Study Visits (V) =	3 (Baseline, Session 5, Session 10)
Effect Size (ES) =	0.09
Power =	0.80
α =	0.05
df (two-way interaction)	4
Cohen's L =	11.94

Subjects/group = [L / (ES * (V-1))] + G = 69

The statistical test used for the primary analysis was not the same method as indicated for the primary analysis when conducting the sample size and power calculations or the same as indicated in the statistical analysis section of the protocol. The study data analyst indicated verbally that they chose a different method based on their expertise and experience that would be more appropriate for the data. The primary analyses were to be completed for the intention to treat group and for protocol completers. Analyses were not conducted for each of these groups.

There are no study stopping rules for efficacy or safety stated, and no planned interim statistical analysis plans. The early unblinding of the study and analysis of the data are not justified per the protocol or for safety reasons. Any analyses presented should be interpreted only as exploratory since they were not conducted per the protocol and do not have sufficient power for the primary hypothesis testing. All analyses conducted should be presented with information of this early termination of the study and unblinding. Data may be used for estimation purposes for future studies and for descriptive purposes, but not recommended for inferential purposes. Decision making

or conclusions should not be drawn from the hypothesis testing conducted and the results interpreted.

There are concerns about the data quality, study conduct and implementation that would also impact the study and any results from this study.

Study Integrity Follow-up:

1. Randomization: Additional information was provided in regards to the randomization implemented for the study:
    - Documentation of the randomization could not be produced.
    - A Block randomization was used and the size was 3. There are only 3 groups, thus the advantage of a block randomization to reduce bias and confounding was not obtained with this strategy.

2. Missing data for the primary endpoint: There was a large amount of data missing for the primary endpoint (CAPS score at analysis timepoints).

Recommendations:

As the study team moves forward with further data review and manuscript preparations, I would recommend that all information regarding the study design, methodology and analyses conducted is outlined clearly in the manuscript(s) in the context of the noted comments as follows:

- The study should be reported with consideration of the Consolidated Standards of Reporting Trials. The 25 item checklist should be followed (http://www.consort-statement.org/checklists/view/32-consort/66-title), in order to allow for the appropriate reporting of how the trial was designed, analyzed, and interpreted; the flow diagram displays the progress of all participants through the trial and should be included (http://www.consort-statement.org/consort-statement/flow-diagram).
- Randomization methodology should include that block randomization was used and that the block size was 3.
- Analyses presented should be interpreted as exploratory since the statistical methods were not specified a priori and conducted per the protocol(s). None of the hypotheses were powered for inferential analysis and interpretation.
- All analyses conducted should be presented with information of early termination of the study and unblinding.
- Data may be used for estimation purposes for future studies and for descriptive purposes, but interpretation of the data is not recommended for inferential purposes.

Comparing Virtual Reality Exposure Therapy (VRET) to Prolonged Exposure (PE) in the Treatment of Soldiers with Posttraumatic Stress Disorder (PTSD)

Overview. This study was a randomized waitlist-controlled clinical trial in which post-Iraq, post-Afghanistan deployed Soldiers with PTSD were randomized to one of three groups: 1) PE (*n* = 54), 2) VRET (*n* = 54), or 3) Waitlist (WL; *n* = 54).

Part 1: Summary of Principal Findings

The two primary hypotheses of the VR/PE study were:

1. PE or virtual reality-enhanced prolonged exposure (VR) would reduce clinical symptoms of PTSD to a greater extent than a waitlist condition.
2. VR therapy would have greater efficacy in PTSD symptom reduction as compared to PE therapy.

In Table 1 below, we present the results of a comparison of PE and VR to the waitlist condition. All groups had a comparable distribution of scores on the CAPS (using the "last week" reference period) at baseline. This was confirmed by the small, non-significant differences in the baseline intercept associated with the PE and VR group assignments. At the treatment midpoint, participants randomized to the PE condition had, on average, a 9.46 point reduction in CAPS scores compared to participants in the waitlist condition which was statistically significant. Similarly, participants in the VR condition had a 5.18 point reduction in CAPS scores, on average, as compared to participants in the waitlist condition. This difference, however, was not statistically significant. By post assessment, participants in the waitlist condition demonstrated, on average, a 10.13 point decrease in CAPS scores that was statistically significant. Both of the treatment arms showed statistically-significant reductions in CAPS scores beyond the waitlist group of 22.43 points for PE and 13.26 points for VR. The conclusion at post assessment is that we reject the null hypothesis of no difference in treatment efficacy between the waitlist condition and either of the active treatment conditions in favor of the alternative hypothesis that both of the active treatments yield a stronger reduction in PTSD symptoms, as measured by the CAPS (week referent) as compared to no treatment.

In Table 2, we present the results of a direct comparison of the efficacy of the PE condition to the VR condition. At baseline, both conditions had similar distributions of CAPS scores. The VR condition had a slightly higher observed mean, but this different did not exceed probabilistic expectation. By post treatment, participants in the PE condition, on average, demonstrated a 32.57 point decrease in CAPS scores. Participants in the VR condition had, on average, a (-32.57 + 9.18 = -23.39) 23.39 point decrease in CAPS scores. This difference was not statistically significant, but suggested that the VR condition was demonstrating less efficacy in score reduction as compared to PE. Over successive time points, the difference in efficacy between PE and VR increased and became statistically significant by the 12-week follow-up assessment. By that point in time, participants in the PE condition evidenced, on average, a 39.72 point reduction in CAPS scores, which was associated with a model-based average score of 38.56. In contrast, participants in the VR condition had a (-39.72 + 15.32 = -24.4) 24.4 point reduction in CAPS scores, on average, which was associated with a model-based average CAPS score of 56.05. Assuming a two-tailed hypothesis about the difference in the efficacy of the PE and VR groups, we have sufficient

evidence to conclude that the null hypothesis of no difference in efficacy can be rejected in favor of an alternative hypothesis that the efficacy of the VR treatment is inferior to the efficacy of the PE treatment.

Part 2: Ethical Issues

Synopsis of Results. At posttreatment and follow-up, VR was less efficacious in reducing CAPS scores than was PE. The lower efficacy associated with VR become statistically significant by 12-week follow-up with PE patients reporting, on average, a 39.71 point reduction on the CAPS as compared with an average 24.40 point reduction for VR patients. The model-based average CAPS score for the PE group was 38.56, whereas the model-based average for the VR group was 56.05 (in the clinically significant range).

Questions for Consideration

1. Table 2a below shows the distribution of patients with CAPS scores ≥ 55 (Schnurr, Friedman, Foy et al., 2003 used CAPS= 45 for PTSD cutoff in veterans; CAPS manual suggests CAPS = 55 is in the "Moderate" range) across the assessment points. As shown, VR is superior to PE at all follow-up assessment points. Is it ethical to randomize PTSD patients to a VR treatment when PE was shown to reduce PTSD to subclinical levels?
2. Is randomization still appropriate, given our data revealed a departure from clinical equipoise (uncertainty about which treatment is superior)?
3. The trial was designed to address whether VR was superior to PE. Given that this hypothesis has been answered, is it ethnical to continue the trial?

Table 2a. Number and proportion of subjects with a CAPS score ≥55 at each time point, by treatment assignment and reference interval of the caps (month [CAPS-Monthly] and week [CAPS-Weekly]

	CAPS-Monthly			CAPS-Weekly		
	PE	VR	Waitlist	PE	VR	Waitlist
Time	n (%)	n (%)	n (%)	n (%)	n (%)	n (%)
Pre-assessment	54 (100.0)	54 (100.0)	54 (100.0)	53 (98.1)	51 (94.4)	50 (92.6)
Midpoint	--	--	--	22 (56.4)	28 (77.8)	40 (76.9)
Postassessment	--	--	--	11 (34.4)	15 (50.0)	36 (68.1)
12-week	9 (33.3)	14 (56.0)	--	8 (29.6)	13 (50.0)	--
26-week	9 (37.5)	8 (47.1)	--	7 (29.2)	9 (52.9)	--

Note: Percentages based on the number of participants providing data at each time point. – indicates that the measure was not collected for the group at that time point.

Table 1. Comparison of CAPS scores, based on "last week" reference, between the prolonged exposure, virtual reality, and waitlist arms during study treatment at baseline, midpoint, and post assessment

Time and parameter	b	Lower 95%	Upper 95%	Model-based means
Baseline				
Intercept (waitlist)	78.89	72.84	84.94	78.89
Prolonged Exposure	-0.61	-9.16	7.94	78.28
Virtual Reality	1.56	-7.00	10.11	80.45
Midpoint				
Intercept	-3.96	-9.65	1.74	74.93
Prolonged Exposure	-9.46	-18.03	-0.89	64.86
Virtual Reality	-5.18	-13.91	3.55	71.31
Post Assessment				
Intercept	-10.13	-16.03	-4.23	68.76
Prolonged Exposure	-22.43	-31.50	-13.36	45.72
Virtual Reality	-13.26	-22.49	-4.04	57.06

Note: Random-intercept linear regression model included all 162 participants (54 randomized per group). Waitlist condition is the referent group.

Table 2.Comparison of CAPS scores, based on "last week" reference, between the prolonged exposure and the virtual reality treatment arms at all assessment times.

Time and parameter	b	Lower 95%	Upper 95%	Model-based means
Baseline				
Intercept	78.28	71.62	84.93	78.28
Virtual Reality	2.17	-7.25	11.58	80.45
Midpoint				
Intercept	-13.41	-20.52	-6.31	64.87
Virtual Reality	4.27	-5.95	14.49	71.31
Post Assessment				
Intercept	-32.57	-40.22	-24.93	45.71
Virtual Reality	9.18	-1.79	20.15	57.06
12-week Follow-up				
Intercept	-39.72	-47.84	-31.59	38.56
Virtual Reality	15.32	3.71	26.92	56.05
26-week Follow-up				
Intercept	-40.14	-48.62	-31.67	38.14
Virtual Reality	13.89	0.99	26.79	54.20

Note: Random-intercept linear regression model included 108 subjects (54 randomized per treatment condition). Prolonged exposure is the referent group.

References

1. Schnurr, P. P, Friedman, M. J., Foy, D. W., et al. (2003). Randomized trial of trauma-focused group therapy for posttraumatic stress disorder: results from a department of veterans affairs cooperative study. Archives of General Psychiatry, 60(5):481-489.

2. Blake, D., Weathers, F., Nagy, L., Kaloupek, D.J. et al. (2000). Clinician-Administered PTSD Scale (CAPS) Instructional Manual. *National Center for Posttraumatic Stress Disorder, Behavioral Science Division, Boston Neurosciences Divisions, West Haven.*